Serial No. 10/042,367
Group Art Unit 2164
Docket No: ARC920000085US1

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

REVISED APPEAL BRIEF – 37 C.F.R § 1.192

U.S. Patent Application 10/042,367 entitled

“AUTOMATED ACCESS TO WEB CONTENT BASED ON LOG ANALYSIS”

Real Party in Interest: International Business Machines Corporation

Related Appeals and Interferences:

None

Status of Claims:

Claims 1-5, 7-11, and 17-19 are pending.

Claim 6, 12-16, and 20-25 are cancelled.

Claims 1-5, 7-11, and 17-19 stand rejected under 35 U.S.C. § 102(e) as being anticipated by U.S. 6,516,312 (Kraft) and are hereby appealed.

Status of Amendments:

No amendments have been submitted after the final rejection.

Summary of Claimed Subject Matter:

(NOTE: All citations are made from the original specification, including the figures.)

The presently claimed invention allows a web crawler to accurately mimic real users, by relying on past user accesses to the Web sites to be crawled. This approach results in a web crawler capable of automatically accessing all the content that a real user would have access to.

The subject matter of independent claim 1 provides a method for determining parameter combinations for automated web crawler (figure 2, element **216**) access to World Wide Web content (page 7, lines 20-21; figure 1, steps **100-106**). A log file (page 7, line 21- page 8, line 6) documents previous real user HTML interactions with said World Wide Web site (Figure 1, element **100**). The method analyzes the log file (page 8, lines 15 and 16; figure 1, step **102**) to determine parameter combinations and to generate synthetic queries (page 8, lines 17-18; figure 1, step **104**) for input to said web crawler. The web crawler then uses the synthetic queries as input for automated access to the World Wide Web content (page 8, lines 18-19; figure 1, step **106**).

The subject matter of independent claim 7 provides a method of increasing web crawler (figure 2, element **216**) penetration of Web databases accessible via HTML forms (page 9). The method includes reviewing previous real user form input data (Figure 1, element **100**), identifying possible HTML form input data for said Web crawler from said previous real user form input data by synthesis of entries for any of: predefined sets, limited text entries or unlimited text entries (page 9, lines 1-3), and providing the identified form input data to the Web crawler (page 8, lines 18-19; figure 1, step **106**) during an instantiation of automated access to said Web databases by said Web crawler.

The subject matter of independent claim 17 provides a method of emulating real user access to World Wide Web content dynamically accessible via an HTML form. The method includes maintaining a log containing real user entries into each input item of said HTML form (page 7, line 21-page 8, line 6), ranking entries for each input item according to their frequency of occurrence (page 9, lines 9-11; page 12, lines 3-10), for each unlimited text entry input item, excluding entries ranked below a predetermined number (page 12, lines 3-6), determining combinations of entries from each set of entries (page 12, line 6), automatically accessing the content using the combinations of entries as HTML input for a web crawler (page 8, lines 18-19; figure 1, step **106**).

Grounds of Rejection to be Reviewed on Appeal:

1. Claims 1-5, 7-11, and 17-19 stand rejected under 35 U.S.C. § 102(e) as being anticipated by U.S. 6,516,312 (Kraft) and are hereby appealed. Was a proper rejection made under 35 U.S.C. § 102(e) with respect to claims 1-5, 7-11, and 17-19 using existing USPTO guidelines?

ARGUMENT:

1. Claims 1-5, 7-11, and 17-19 stand rejected under 35 U.S.C. § 102(e) as being anticipated by U.S. 6,516,312 (Kraft) and are hereby appealed. Was a proper rejection made under 35 U.S.C. § 102(e) with respect to claims 1-5, 7-11, and 17-19 using existing USPTO guidelines?

To be properly rejected under 35 U.S.C § 102(e), each and every claim element must be shown in a single reference or its inherency set forth.

Applicants will set out many specific arguments below to distinguish over the Kraft reference used in the rejection. However, it is important to recognize at least two very basic distinctions between Kraft and the present invention.

First, Kraft is prevented from employing the claimed steps because Kraft generates all secondary searches from a "Result Set keyword", not the presently claimed "log file containing user queries". In other words, the present invention uses the real world questions (queries) asked during a database search (using HTML forms) to generate future synthetic queries, while Kraft analyzes the answers (results) of a search and then proceeds in a divergent path. By the examiner's own admission, see Final office action dated 6/12/2006, page 5, line 2, " ...none of the independent claims detail the nature of any search results… Yet, the examiner then equates the "search results" of Kraft to the present invention claims. This constitutes a clear error in application of the prior art to the claim elements. When the examiner is pointing to figures 6A and 6B, he is not pointing to a log file containing user queries at all, only search result sets generated by queries (title of Kraft's figure 6A is SEARCH RESULTS). Additional proof of this characterization can be found throughout Kraft. For example, col.1, lines 12-19 cite "....this invention pertains to a computer software product for dynamically associating keywords encountered in abstracts or summaries of a search result set....". Kraft uses the search results to match keywords with existing "dictionary terms" in a local database to give additional info to the user. Claim 1, specifically claims a "log file containing user queries"; claim 7 specifically claims "synthesis of entries"; and claim 17 specifically claims "a log containing real user entries". None of these claims is directed to an evaluation of search results as is taught by Kraft.

Second, the examiner has inaccurately equated the Kraft terms "search results" and "web browser" throughout his rejections to the claimed respective terms "log files" and "web crawler". These terms are well known in the art, and further described in the specification, and should not be misused. Surely, one skilled in the art would not confuse the "results of a search" with a "log file". Also, a PC's "web browser" which constitutes a well known user interface where a www link can be entered should not be confused with a web crawler operated separately from the user's web browser which performs the functions of the search itself.

These very basic differences, as well as many other differing claimed elements, prevent Kraft from anticipating the presently claimed invention. Without a teaching of using a log file containing user queries and/or an automated web crawler, Kraft cannot even satisfy the minimum required claim elements.

Applicants wish to emphasize that both the pending patent application and the primary reference (Kraft et al.) are commonly assigned and, at the time the claimed invention was made, were both subject to an obligation to be assigned to IBM. It will be shown below that the Kraft reference does not provide many of the elements of the claims and therefore cannot be properly rejected under 35 U.S.C. §102(e). A shift to a 35 U.S.C. §103 rejection would result in disqualification of this reference as prior art.

The examiner has rejected claims 1-5, 7-11, and 17-19 under 35 U.S.C. §102(e) as being anticipated by Kraft et al. (USP 6,516,312). To be properly rejected under 35 U.S.C. §102, each and every element of claims must be disclosed in a single cited reference. The applicants, however, contend that the presently claimed invention cannot be anticipated in view of the '312 reference.

The Kraft et al reference (hereafter Kraft) is primarily cited for its provision of new search queries generated from a domain-specific user query that was dynamically associated with keywords. The Kraft reference teaches away from the present invention by generating a <u>new search result</u> from a set of previously prepared abstracts and by providing additional, supplemental information to each user query. Since a search engine repository is updated with this additional information, subsequent executions of the same user query will not, and are not intended to generate the same or equivalent search results, but rather provide new, different information to the user.

With regard to independent claims 1, 7, and 17, the examiner has cited figures 6A and 6B of Kraft to equate to a log file containing previous user queries. First, figures 6A and 6B are not log files, but rather search result listings (see title of figure – SEARCH RESULTS). As made clear above, figures 6A and 6B are simply search result sets with keywords noted. These keywords are matched to a local dictionary of terms (see col. 8, lines 35-38, etc.) – domain-specific dictionary 110.

By contrast, the present invention discloses an ordered set of parameters in a log file chosen such that automated access (as opposed to Krafts' user selecting of highlighted keywords for more info; see col. 11, lines 29-30, which state "the user, desiring to learn more about a desired term RMI, selects this term…", emphasis added) to the same WWW content as would be accessed manually, by a real user, is provided. Each parameter stored in a log file of the present invention is comprised of a name and associated value, specifically, an input field name in a WWW form and an associated input value to this field. In other words, the presently claimed invention seeks to reverse engineer a manual access of web content by automatically answering a question (i.e. input field name) presented by a web site with an answer (i.e. input value) that is based on a stored set of user responses (i.e. parameters values) to the same question (i.e. parameter name) presented by the same WWW form. A combination, as specified by the present invention, is a set of parameters that are individually input to a web form, whereas the combination disclosed by Kraft is number of distinct URLs and keywords combined to create a single query string.

Applicants contend that abstracts contained in the log file maintained in search service provider as characterized by the examiner cannot be used to determine parameter combinations, nor can they be used in attaining access to web content, wherein access is automated or otherwise.

Additionally, in the Final office action dated 6/12/06 (page 4), the examiner cites that the claims fail to provide for a "query log" and therefore the arguments are deemed moot. While the arguments to the rejection contained the abbreviated citation "query log", all independent claims provide for a "log file containing user queries". The functionality and use are the same, further described in the specification, and provided in the original claims and drawings. The examiner cannot dismiss the argument on one hand and then not consider the argument in light of the actual language and its functionality.

With respect to dependent claims 2, 3, and 6, the examiner has cited figure 6A as illustrating parameters and ranking. First, figure 6A illustrates a listing of search results. No teaching has been provided by the examiner directed to ranking query entries themselves. To simply show ranking, which is well known, without a teaching of ranking the same elements, leaves the argument without merit. The examiner must not only show ranking, but also that the ranking provides the same purpose or function - ranking of entries to queries.

The examiner appears to have equated parameters disclosed and claimed in the present invention with a text string and arbitrary URLs containing the same text string. Such a text string, for example, "RMI" in figure 6A as cited by the examiner, is not a parameter but rather, a simple keyword. A parameter of the present invention requires, for example, both a name (e.g. "zip code") and an associated value (e.g. "95120") appropriate for the name. In order for a web crawler to gain automated access to certain web content, the present invention teaches the determination of a value for a name component of a given parameter requested by a web site, for example a value for a zip code. In essence, when a Web site presents the question "What is the zip code?", a crawler reads previous responses stored for this question, answer the question with a value or values, "95120”, based on what it read.

Because keywords and URLs are not parameters (i.e. they are not input fields with appropriately specified input values), their combination cannot be used to appropriately fill out HTML forms having fields requiring input. Additionally, keywords cited in figure 6a of the Kraft reference are not parameters because there are no input fields requiring input values.

With regard to dependent claims 4, 10, 11, and 18 the examiner has cited figure 6A of Kraft as suggesting both limit and unlimited text entries with removal of stop words and stemming. First, the examiner cites that he has noted certain stop words “by”, “and” and “the” not within the search results. This argument bears no weight as no teaching of removing stop words has been made. To simply state that a specific chosen few stop words are not present does not equate to a removal step. In fact, the examiner has specifically chosen to ignore included stop words such as “with, and “or”. Clearly stop words have not been removed. Secondly, pointing to an abbreviated term such as “monthly publication and author’s full name” does not equate to “stemming remaining words”. In fact, the term “programmer” is not stemmed. Clearly, the remaining terms are not stemmed.

With regard to dependent claims 5, 9, and 19, figure 3 of the Kraft reference is also cited by the examiner as suggesting a proxy server used in the description of the present invention. A proxy server of the present invention refers to a computer or program that is transparent to a client; a client does not see or know that a proxy server exists. Instead, a client sees web content produced by a web server to which the client is connected. A proxy server records communication between a web server and client silently and transparently (i.e. without requiring a client to know of its existence). In contrast, the search service provider pointed to by the examiner in the Kraft reference is, in fact, a web server directly providing content in response to a client’s request. It is implied, therefore, that a client is aware of the search service provider’s existence by the fact that a client issues a request for content directly to the search service provider. Thus, a search service provider cannot be a proxy server for the following reasons: it provides content; a client is aware of a direct connection to it; and it does not intercept communications, transparently or otherwise.

As per claim 8, the examiner has equated storing annotated abstracts in a local database with maintaining a log file. However, no explicit recitation of a log file exists with Kraft.

SUMMARY

As has been detailed above, the Kraft et al reference does not provide for the specific claimed details of applicants' presently claimed invention and therefore a rejection under 35 U.S.C. § 102(e) is deemed improper. It is believed that this case is in condition for allowance and reconsideration thereof and early issuance is respectfully requested.

As this Appeal Brief has been timely filed within the set period of response, no request for extension of time or associated fee is required. However, the Commissioner is hereby authorized to charge any deficiencies in the fees provided, to include an extension of time, to Deposit Account No. 50-4098.

Respectfully submitted by
Applicant's Representative,

/ramraj soundararajan/

Ramraj Soundararajan
Reg. No. 53,832

IP Authority, LLC
9435 Lorton Market Street #801
Lorton, VA 22079
(571) 642-0033

February 21, 2007

Claims Appendix:

1. (Previously Presented) A method of determining parameter combinations for automated web crawler access to World Wide Web content that is accessible based on parameters resulting from real user interactions with a World Wide Web site, said method comprising:

maintaining at least one log file containing user queries resulting from previous real user HTML interactions with said World Wide Web site;

analyzing said log file to determine parameter combinations and to generate synthetic queries for input to said web crawler, said web crawler using said input for automated access to said World Wide Web content.

2. (Previously Presented) A method of determining parameter combinations for automated access to World Wide Web content that is accessible based on parameters resulting from real user interactions with a World Wide Web site, as per claim 1, said user queries comprising entries, said analyzing step further comprising

ranking entries according to their frequency of occurrence;

for a set of entries resulting from unlimited text entries, excluding entries ranked below a predetermined number; and

wherein said synthetic queries are determined by producing combinations of entries from each set of entries.

3. (Previously Presented) A method of determining parameter combinations for automated access to World Wide Web content that is accessible based on parameters resulting from real

user interactions with a World Wide Web site, as per claim 2, wherein said synthetic queries are determined by producing all combinations of entries from each set of entries.

4. (Original) A method of determining parameter combinations for automated access to World Wide Web content that is accessible based on parameters resulting from real user interactions with a World Wide Web site, as per claim 2, wherein entries resulting from limited text entries and unlimited text entries have stop words removed and remaining words stemmed.

5. (Original) A method of determining parameter combinations for automated access to World Wide Web content that is accessible based on parameters resulting from real user interactions with a World Wide Web site, as per claim 1, wherein said log file is maintained by a proxy server that logs communications between a client and a Web server resulting from real user accesses to said World Wide Web content.

6. (Canceled)

7. (Previously Presented) A method of increasing web crawler penetration of Web databases accessible via HTML forms, said method comprising:

reviewing previous real user form input data;

identifying possible HTML form input data for said Web crawler from said previous real user form input data by synthesis of entries for any of: predefined sets, limited text entries or unlimited text entries; and

providing said identified form input data to said Web crawler during an instantiation of automated access to said Web databases by said Web crawler.

8. (Previously Presented) A method of increasing web crawler penetration of Web databases accessible via HTML forms, as per claim 7, wherein said previous form input data are maintained in a log file.

9. (Original) A method of increasing web crawler penetration of Web databases accessible via HTML forms, as per claim 8, wherein said log file is maintained by a proxy server.

10. (Original) A method of increasing web crawler penetration of Web databases accessible via HTML forms, as per claim 7, wherein said synthesis comprises:

ranking any entries for predetermined sets;

ranking any entries for limited text entries;

ranking any entries for unlimited text entries;

excluding entries for unlimited text entries ranked below a predetermined number; and

pairing entries from each set of ranked entries.

11. (Original) A method of increasing web crawler penetration of Web databases accessible via HTML forms, as per claim 10, wherein said synthesis further comprises:

removing stop words and stemming remaining words for entries resulting from limited text entries and unlimited text entries.

12 – 16 (Canceled)

17. (Previously Presented) A method of emulating real user access to World Wide Web content dynamically accessible via an HTML form, said method comprising:

maintaining a log containing real user entries into each input item of said HTML form;

ranking entries for each input item according to their frequency of occurrence;

for each unlimited text entry input item, excluding entries ranked below a predetermined number;

determining combinations of entries from each set of entries; and

automatically accessing said content using said combinations of entries as HTML input for a webcrawler.

18. (Original) A method of emulating real user access to World Wide Web content dynamically accessible via an HTML form, as per claim 17, wherein entries resulting from limited text entries and unlimited text entries have stop words removed and remaining words stemmed.

19. (Original) A method of emulating real user access to World Wide Web content dynamically accessible via an HTML form, as per claim 17, wherein said log file is maintained by a proxy server that logs communications between a client and a Web server resulting from real user accesses to said World Wide Web content.

20 – 25 (Canceled)

Evidence Appendix

None

Related Proceedings Appendix

None